Self-location and causal context

Friederich, Simon

Published in:
Grazer Philosophische Studien

DOI:
10.1163/18756735-09302008

Publication date:
2016

Self-location and causal context

Simon Friederich

email@simonfriederich.eu

University College Groningen, Hoendiepskade 23/24, NL-9718 BG Groningen and
Faculty of Philosophy, University of Groningen, Oude Boteringestraat 52, NL-9712 GL Groningen, The Netherlands

Abstract: The paper proposes a novel principle of rational self-locating belief that refers to the epistemic agent's causal context. The principle is motivated and applied to some of the most-discussed problems of self-locating belief including the Doomsday Argument, the Serpent's Advice scenario, the Presumptuous Philosopher problem, the Sleeping Beauty problem, and the problem of confirmation in the Everett interpretation. It is shown to yield plausible verdicts in all these cases.

1 Introduction

In a typical problem of self-locating belief, observers figure out their rational credences as regards what the world is like by properly taking into account the available self-locating evidence, i.e. the available evidence as to where and when they exist, and as who among all observers. Whether some bit of self-locating information is available to an agent can have considerable impact on her rational credences concerning matters non-self-locating. Problems of self-locating belief have received increasing attention by philosophers in recent years. The most widely discussed such problems include the Doomsday Argument (Gott 1993, Leslie 1996), the Presumptuous Philosopher problem (Bostrom 2002a), the Sleeping Beauty problem (Elga 2000), and the problem of confirmation in the Everett interpretation of quantum mechanics (Lewis 2007). (For overviews and comparative assessments of different versions of these problems, see, for example, (Bostrom 2002a) and (Bradley 2012) and references therein. All these problems are reviewed and discussed in the present paper.)

The present paper adopts the following set-up for problems of self-locating belief: in the first step, a set of hypotheses as to how the world might be is chosen and probabilities $Pr()$ are assigned to them. These can be relative frequencies, objective chances or other probabilities arrived at on grounds of systematic considerations (examples will follow below). In what follows, I refer to the probabilities $Pr()$ as *non-anthropic input probabilities*: contrary to what Carter's *anthropic principle* (Carter 1974) reminds us to

do, they do not yet take into account any constraints on possible observations that arise from the fact that what is observed must be compatible with the existence of observers. In the second step, the situation of some (hypothetical) epistemic agent is considered who either has or lacks some bit of self-locating information that might be relevant to her rational assessment of the hypotheses at issue; finally, in the third step, it is asked whether the agent's rational credence $cr(A)$ with respect to some (non-self-locating) proposition A is the same as the non-anthropic input probability $Pr(A)$ and, if not, how it differs. In other words, in problems of self-locating belief as set up here the challenge is to compute an epistemic agent's rational credence $cr(A)$, given the amount of self-locating information that the agent has, from the non-anthropic input probability $Pr(A)$. As we will see, in many examples this challenge is far from trivial.

What makes problems of self-locating belief so fascinating is that any suggested prescription for connecting the non-anthropic input probabilities and the rational credences that appears to work well for some problems seems to run into trouble for others. For instance, the so-called *self-indication assumption* (SIA) (Bostrom 2002a, 66), which can be interpreted as a prescription of how to link input probabilities and rational credences, delivers an intuitively plausible verdict on the Doomsday Argument, but it also delivers a catastrophically implausible verdict on the Presumptuous Philosopher problem. (Both are explained in detail below.)

The main idea of the present approach to connect the non-anthropic input probabilities and the rational credences is the following: typically, observers can take over the non-anthropic input probabilities as their rational credences by identifying the two, but exceptions occur whenever observers are disoriented about their place in the causal order of observers, in which cases they have to correct for effects that result from their being so disoriented. I propose a causal principle of self-locating belief that makes this idea more specific and precise. The results obtained from this principle for the problems mentioned are: (i) that the Doomsday Argument is invalid, (ii) that the presumptuous philosopher is presumptuous (i.e. his reasoning to be rejected), (iii) that the thirder view is correct in the standard version of Sleeping Beauty, and (iv) that the Everett interpretation does not receive automatic confirmation by arbitrary data.

Unlike the present work, most recent systematic work on rational self-locating belief adopts a diachronic perspective that develops and evaluates updating policies for rational credences in the light of newly won (or lost) self-locating evidence. (For examples of suggested policies, see (Bradley 2011, Briggs 2010, Cozic 2011, Kim 2009, Meacham 2008, 2010, Moss 2012, Schulz 2010, Schwarz 2012, 2015, Stalnaker 2008, Titelbaum 2008, 2013).) The present work differs in that it focuses on the relation between the non-anthropic input probabilities and the rational credences, which is *not* that between prior and posterior credences given newly acquired (or lost) self-

locating evidence. Working with non-anthropic input probabilities has the advantages that, first, no *prior* credences are needed from which to compute the looked-for posteriors and that, second, the question can be set aside which observer stages (possible world "centres", in the terminology to be used) are successive stages of one and the same observer. At the same time, working with non-anthropic input probabilities has the drawback that uniquely determined non-anthropic input probabilities may not be available or that there may be several candidate relevant non-anthropic input probabilities between which it is difficult to choose. Given these relative advantages and drawbacks, the present approach should be seen not as rival to diachronic approaches but as complementary, having its own comparative assets and problems, which may be more or less pronounced depending on the scenario under consideration.

The structure of the remaining sections of this paper is as follows: Section 2 presents the formal basis of the approach to be used. Sections 3 and 4 introduce two simple (conflicting) candidate principles of how to obtain the rational credences from the non-anthropic input probabilities. Both are shown to correspond to internally coherent styles of reasoning that may be applied to various problems of self-locating belief (roughly speaking, the first to reasoning that avoids the SIA, the second to reasoning that uses it), both are also shown to have highly counterintuitive consequences. In view of these counterintuitive consequences and in response to them, Section 5 motivates and proposes the above-mentioned causal principle of how to obtain the rational credences from the non-anthropic input probabilities. Section 6 applies this principle to the Doomsday Argument and the Presumptuous Philosopher problem and shows that it yields plausible recommendations for them, Section 7 applies it to the Sleeping Beauty problem, Section 8 to the Everett interpretation of quantum mechanics. Finally, Section 9 closes the paper with an outlook on challenges that remain.

2 The formalism

The formalism to be used here is a standard Lewisian one that identifies propositions with sets of possible worlds. Uncentred and centred propositions are distinguished. An uncentred proposition p (scare quotes are omitted for the sake of easy readability) corresponds to the collection of (uncentred) possible worlds V, W, X, ..., in which it is true, and the uncentred worlds themselves are equivalent to the propositions that describe them completely. For a proposition p that corresponds to the collection V, W, X, ... of worlds, p is represented as the disjunction $V \vee W \vee X \vee \ldots$.

Many uncentred possible worlds harbour observers, and these have their own individual perspectives within the worlds they live in. It is common to refer to the distinct points of perspective inside some world as its *centres*.

There is no fixed and determinate criterion of what counts as a centre. Typically what one takes to be a centre depends on the problem and context at issue. For many purposes, one identifies the centres of a world with the conscious observers who inhabit it, but sometimes a more fine-grained distinction into *observer-moments* (Bostrom 2002a), i.e. stages of one and the same observer at different times, is preferable. In the problem cases to be discussed, there are well-established conventions as to how to individuate centres as observers or observer-moments. I adopt these conventions for convenience, but it should be kept in mind that the choice of centres has nontrivial implications and is often a decisive move in setting up a problem of self-locating belief. (The challenge of choosing an appropriate class of centres in a possible world is the more generic reference class problem's manifestation in the theory of rational self-locating belief.)

A *centred* possible world V_i corresponds to an uncentred world V with one centre picked out and denoted by the variable i. Centred propositions correspond to collections of centred possible worlds, which means that they can be represented in the form $V_i \vee W_j \vee X_k \vee ...$ (where any uncentred world V can appear with different indices several times).

Following David Lewis (see Lewis 1979 and, for an application to a problem of self-locating belief, Lewis 2001), I allow that uncentred possible worlds may be decomposed into maximally unspecific centred possible worlds, i.e. worlds V are equated with maximal disjunctions of centres V_i in the form $V = V_1 \vee V_2 \vee ...$, where V_1, V_2, ... are the centres of the world V. Only those uncentred worlds that contain some centre (or centres) can be decomposed into centred ones.

3 A simple credence-prior link

Self-locating evidence is evidence as to which of the centres V_1, V_2, ... in some uncentred world V one might be. Let us start by asking what the rational credence of being the centre V_{i_j} is, *given* that one is one of the centres V_{i_1}, V_{i_2}, ... V_{i_n}. To tackle this question, let us assume that there is a credence function $cr(\,)$ from which the rational credence of being V_{i_j}, given that one is either V_{i_1} or V_{i_2} or ... or V_{i_n} (while not knowing which), is obtained as the conditional probability $cr(V_{i_j}|V_{i_1} \vee V_{i_2} \vee ... \vee V_{i_n})$. A straightforward, and in many cases plausible, answer is that one should be *indifferent* as to which of the centres V_{i_j} on the right hand side of the "|" one is, i.e. one should have equal credence in being any of these centres. This suggestion is expressed in the following formal principle IND (where "IND" stands for "indifference"):

$$\text{(IND)} \qquad cr(V_{i_j}|V_{i_1} \vee V_{i_2} \vee ... \vee V_{i_n}) = \tfrac{1}{n}$$

where i_j is one of the $i_1, ..., i_n$.

Indifference principles in the spirit of IND are defended by various authors, for example by Vilenkin, Bostrom and Elga and applied in a variety of contexts where self-locating evidence matters. Vilenkin offers his *principle of mediocrity* (Vilenkin 1995) as a principle of typicality in cosmological contexts; Bostrom strives for more generality with his *self-sampling assumption* (Bostrom 2001, 2002a,b) according to which "[o]ne should reason as if one were a random sample from the set of all observers in one's reference class", leaving open the reference class to be used (Bostrom 2002a, 57); finally, Elga (Elga 2004) defends "indifference" as a weak form of IND, which applies only if the psychological states of the V_{i_j} are subjectively indistinguishable.

Even the weakest of the indifference principles—Elga's—has been subjected to a sustained critique in the literature (for example in (Weatherson 2005) and (Schwarz 2015, Section 7)), and there is little reason to suppose that IND is always attractive: there are certainly circumstances where one's evidence leaves it open which centre among several one is and where it is rational not to be indifferent as to which one it is. Nevertheless, I suggest that IND be accepted for the purposes of this paper as a working assumption. There are at least three reason why this seems to be a good idea: first, IND delivers plausible consequences in some simple scenarios such as Bostrom's *dungeon* (Bostrom 2002a, 59f., see also his discussion of two scenarios due to Leslie in Bostrom 2002a, 62f.), and it is not easy to come up with a similarly powerful, less controversial principle to recover these consequences while rejecting IND; second, in the examples to be discussed, as will be pointed out, deviations from IND would have to be radical in order to lead to verdicts that differ qualitatively in interesting ways; and, third, due to its simplicity IND is worthy of exploration in itself, even if it turns out to be nothing more than a first-order approximation to a more general and more satisfying principle.

Applicability of IND is not confined to situations where one is certain that the actual uncentred world is V. Using $cr(V_i|V \vee W \vee X \vee ...) = cr(V_i|V) \cdot cr(V|V \vee W \vee X \vee ...)$ we obtain from IND:

$$\text{(GenIND)} \qquad cr(V_i|V \vee W \vee X \vee ...) = \tfrac{1}{N_V}\, cr(V|V \vee W \vee X \vee ...)\,,$$

where N_V is the total numbers of centres in V. As demonstrated by Vilenkin, Bostrom, Elga and many others, GenIND can be used to derive very interesting results concerning rational self-locating belief.

As announced in the introduction, the present paper uses a set-up for problems of self-locating belief that is based on what I call *non-anthropic input probabilities* $Pr(\,)$, assigned to the uncentred worlds V , W, X, ... Typically, these probabilities are either objective chances (such as those of quantum events or of outcomes of coin tosses and other random events) or (suitably idealised) relative frequencies. As it turns out, in all the scenarios to be discussed here there are some probabilities that suggest themselves for

being treated as the non-anthropic input probabilities. Examples of non-anthropic input probabilities to be used are: a healthy couple's probability of conceiving a child when having intercourse, the probability that some asteroid with well-defined boundary conditions will collide with the earth and extinguish human life, the probability of a fair coin to come up $Heads$, and, in the final example to be discussed, an agent's prior degree of belief in the Everett interpretation as based on philosophical reasoning while ignoring empirical data.

In all the scenarios to be discussed, the main challenge is to determine the rational credences from the non-anthropic input probabilities by adequately considering the available self-locating evidence. One way to approach this challenge is by asking under which conditions the non-anthropic input probabilities just *are* the rational credences, given the available self-locating evidence. The most straightforward possible answer to this question that I discuss here is that the non-anthropic input probabilities directly translate into the rational credences when one's self-locating information is *maximally unspecific* in that

$$\text{(MU)} \qquad cr(V|V \vee W \vee X \vee ...) = Pr(V|V \vee W \vee X \vee ...)\,,$$

where "MU" stands for "maximally unspecific". According to MU, if all that one knows about one's location is that one might just be any of the centres in the worlds V, W, X, ..., then one's rational credence with respect to V is the non-anthropic prior $Pr(V|V \vee W \vee X \vee ...)$ itself.

Though simple and apparently natural, the principle MU has some very counterintuitive consequences. In particular, in combination with GenIND it gives rise to the notorious Doomsday Argument in its different versions. Let us review a simple version of this argument: consider two hypotheses H_1 and H_2 as to how many observers are ultimately going to have existed (N_1 according to H_1, N_2 according to H_2). To simplify matters, let us focus on human observers and assume that they alone are the elements of our reference class (see Bostrom 2001, note 1, for considerations on what happens if we drop this assumption). Next, assume that you are the n-th observer ever to exist, where $n < N_1$ and $n < N_2$, i.e. you are certain to be either $H_{1,n}$ or $H_{2,n}$. Finally, let us assume that $N_1 \ll N_2$ in that H_1 predicts that a fatal asteroid strike will lead to an untimely end of humanity, so that much less human beings will have lived according to H_1 than H_2. Non-anthropic input probabilities $Pr(H_1)$ and $Pr(H_2)$ are assigned to H_1 and H_2 based on information about the relative frequencies of asteroid strike in some (hopefully) well-chosen reference class of asteroids.

Given these assumptions, what are your rational credences concerning H_1 and H_2? Assuming that H_1 and H_2 exhaust the space of possibilities to be considered (in that $cr(H_1) = cr(H_1|H_1 \vee H_2)$ and $cr(H_2) = cr(H_2|H_1 \vee$

H_2)) we obtain

$$\begin{aligned}
\frac{cr(H_1|H_{1,n} \vee H_{2,n})}{cr(H_2|H_{1,n} \vee H_{2,n})} &= \frac{cr(H_{1,n} \vee H_{2,n}|H_1)\, cr(H_1)}{cr(H_{1,n} \vee H_{2,n}|H_2)\, cr(H_2)} \\
&= \frac{cr(H_{1,n}|H_1 \vee H_2)}{cr(H_{2,n}|H_1 \vee H_2)} \\
&= \frac{N_2}{N_1} \cdot \frac{cr(H_1)}{cr(H_2)} \\
&= \frac{N_2}{N_1} \cdot \frac{Pr(H_1)}{Pr(H_2)}\,, \qquad (1)
\end{aligned}$$

where Bayes' theorem has been used in the first line, the definition of conditional probability in the transition from the first to the second, GenIND in the transition from the second to the third, and MU in the transition from the third to the fourth.

Thus, based on MU and GenIND we obtain the result that your relative rational credences with respect to the hypotheses H_1 and H_2 are inversely proportional to the numbers of observers N_1 and N_2 that exist according to them. A smaller total number of observers ever to have lived is confirmed over a larger one. The effect persists if more hypotheses H_3, H_4, ... that other numbers of observers are considered. It also persists if modest deviations from GenIND are allowed or the reference class of observers (i.e. centred worlds) is modestly altered. Furthermore, for any choice of reference class an analogously structured argument can be constructed with suitable hypotheses H_1, H_2. Some version of the Doomsday Argument looms for any choice of reference class and hypotheses considered.

However, the conclusion of the Doomsday Argument is implausible, and wildly differing diagnoses have been offered in the literature as to what might be wrong with it.[1] To determine what exactly makes it so bizarre it is useful to consider one more scenario that is similar to the Doomsday Argument but highlights the counterintuitive consequences of MU in combination with GenIND in an even more drastic manner. The example is due to Bostrom, who calls it "Serpent's Advice":

> Eve and Adam, the first two humans, knew that if they gratified their flesh, Eve might bear a child, and if she did, they would be

[1] To name just a few examples of reactions to the Doomsday Argument, Norton 2010 challenges the Bayesian reasoning based on which the conclusion is derived; Neal 2006 and Bostrom 2001, along different lines, hold that the problem is an artifact of an inappropriate choice of reference class, while others, notably Pisaturo 2009, Lewis 2010 and Bradley 2012, argue that the conclusion of the Doomsday Argument (essentially the last line of Eq. (1)) is sound and only apparently unacceptable. Dieks, in contrast, argues for rejecting the Doomsday Argument along lines as if the self-indication assumption SIA, discussed here in the Section 4, were valid (Dieks 2007), without committing himself to this assumption. The considerations presented here are in line with Dieks' account. The principle to be advanced in Section 5 gives a systematic justification of why this does not commit one to the SIA.

> expelled from Eden and would go on to spawn billions of progeny that would cover the Earth with misery. One day a serpent approached the couple and spoke thus: "Pssst! If you embrace each other, then either Eve will have a child or she won't. If she has a child then you will have been among the first two out of billions of people. Your conditional probability of having such early positions in the human species given this hypothesis is extremely small. If, one the other hand, Eve doesn't become pregnant then the conditional probability, given this, of you being among the first two humans is equal to one. By Bayes's theorem, the risk that she will have a child is less than one in a billion. Go forth, indulge, and worry not about the consequences!" (Bostrom 2001, 366)

The serpent's reasoning is correct, given MU and GenIND, in complete analogy with the Doomsday Argument. To see this, suppose that Adam and Eve consider only two hypotheses (possible worlds): one, H_1, according to which they remain the only two humans ever to have lived, and one, H_2, according to which $N > 10^9$ humans are ultimately going to have lived. Adam's self-locating evidence tells him that he is the first observer ever to have existed, i.e. either $H_{1,1}$ or $H_{2,1}$. Using this information, he can compute the ratio between his rational credences with respect to H_1 and H_2 in analogy to Eq. (1) as:

$$\frac{cr(H_1|H_{1,n} \vee H_{2,n})}{cr(H_2|H_{1,n} \vee H_{2,n})} = \frac{N}{2} \frac{Pr(H_1)}{Pr(H_2)} . \tag{2}$$

Let us assume, for the sake of simplicity, that the probability of getting pregnant after intercourse is approximately 1/2, which translates into $Pr(H_1) \approx Pr(H_2) \approx 1/2$. Accordingly, $Pr(H_1)$ and $Pr(H_2)$ approximately cancel out and the result of Eq. (2) becomes $N/2 \approx 10^9$. Thus, if we grant the serpent the use of MU and GenIND, we have to concur that Adam and Eve should consider the pregnany risk as completely negligible (or, perhaps to their chagrin, their abilities to have children as weirdly reduced).

The main source of the counterintuitive character of the results expressed in Eqs. (1) and (2) is the factor $\frac{N_2}{N_1}$, in the third line of Eq. (1) and $\frac{N}{2}$ in Eq. (2). To eliminate it, one may either look for an alternative to GenIND or for one to MU. As already announced, I will keep GenIND as a working assumption and look for an alternative to MU. In the following section I discuss the prospects for identifying rational credences with the non-anthropic priors given *maximally specific* self-locating evidence.

4 An alternative principle

In this section I discuss a straightforward alternative to MU that elegantly avoids the Doomsday Argument and the serpent's advice. Unfortunately, as it turns out, this alternative has highly counterintuitive consequences on its own. In fact, as we shall see, it inherits the implausibility of the disastrous *self-indication assumption* (SIA) (Bostrom 2002a, 66).

The alternative to MU is that the non-anthropic prior translates into the rational credence conditional on *maximally specific* ("MS") self-locating evidence:

$$\text{(MS)} \qquad cr(V|V_i \vee W_j \vee X_k \vee ...) = Pr(V|V \vee W \vee X \vee ...)$$

for arbitrary centred possible worlds V_i, W_j, X_k, ... The disjunction $V_i \vee W_j \vee X_k \vee ...$ is to be understood as containing exactly one (arbitrary) centred world V_i for each uncentred world V and, for each uncentred world that has no centre, that uncentred world itself. Thus, according to MS the non-anthropic prior $Pr(V)$ translates into the credence that one rationally ought to have with respect to V in a situation where, for each world V, W, X, ... that might be the actual world, one can exclude all but one centre V_i, W_j, X_k, ... as one's own ("actual" one).

Considered by itself, the principle MS is implausible: why should the non-anthropic prior—which one might characterise as the probability arrived at by *ignoring* the self-locating evidence—give the rational credence in a situation when one has *maximally specific* self-locating information? Nevertheless, it is useful to consider the ramifications of MS as it provides the most straightforward way to get rid of the implausible conclusions of the Doomsday Argument and the Serpent's Advice scenario. Appealing to MS, we directly obtain the desired result:

$$\frac{cr(H_1|H_{1,n} \vee H_{2,n})}{cr(H_2|H_{1,n} \vee H_{2,n})} = \frac{Pr(H_1)}{Pr(H_2)}\,. \tag{3}$$

This means that the ratio of the rational credences when having the self-locating evidence "I am the n-th observer" (expressed as "$H_{1,n} \vee H_{2,n}$") is precisely that of the non-anthropic priors, just as one would have expected.

However, besides being implausible in itself, the principle MS has the unwelcome feature that, as already remarked, its consequences are the same as those of the self-indication assumption in that it states that the rational credence in some hypothesis is proportional to the number of observers that exist according to it. This can be seen as follows (where I use $cr(V)$ as a

shorthand for $cr(V|V \vee W \vee X \vee ...)$ and similarly for $Pr(V)$):

$$\begin{aligned} Pr(V) &= cr(V|V_i \vee W_j \vee X_k \vee ...) \\ &= \frac{cr(V_i)}{cr(V_i) + cr(W_j) + cr(X_k) + ...} \\ &= \frac{1/N_V \, cr(V)}{1/N_V \, cr(V) + 1/N_W \, cr(W) + 1/N_X \, cr(X) + ...}\,, \end{aligned} \tag{4}$$

where the first line uses MS, the second line uses the definition of conditional probability, and the third line uses GenIND (where N_V is the number of centres in V and analogously for N_W, N_X etc.).

Evaluating Eq. (4) for two uncentred worlds V and W one obtains

$$\frac{cr(V)}{cr(W)} = \frac{N_V}{N_W} \cdot \frac{Pr(V)}{Pr(W)}\,, \tag{5}$$

which implies, implausibly, that one's rational degree of belief in a possible world is proportional to the number of observers that exist in it. This is equivalent in its consequences with Bostrom's SIA. (Bostrom does not rely on the distinction between rational credences and non-anthropic priors as the present paper does and, accordingly, formulates the SIA somewhat differently, as an assumption about rational priors. For practical purposes Eq. (4) and the SIA are equivalent.)

The SIA's unacceptable consequences are highlighted especially pointedly in the *Presumptuous Philosopher* scenario:

> It is the year 2100 and physicists have narrowed down the search for a theory of everything to only two remaining plausible candidate theories, T_1 and T_2 (using considerations from super-duper symmetry). According to T_1 the world is very, very big but finite and there are a total of a trillion trillion observers in the cosmos. According to T_2 , the world is very, very, very big but finite and there are a trillion trillion trillion observers. The super-duper symmetry considerations are indifferent between these two theories. Physicists are preparing a simple experiment that will falsify one of the theories. Enter the presumptuous philosopher: "Hey guys, it is completely unnecessary for you to do the experiment, because I can already show to you that T_2 is about a trillion times more likely to be true than T_1 !" (Whereupon the presumptuous philosopher explains the Self-Indication Assumption.) (Bostrom 2002a, 124)

The philosopher's claim seems indeed presumptuous, and following his recommendation would clearly be irrational.[2] Generally speaking, given non-anthropic priors of similar size for two hypotheses, it does not seem rational

[2]See, however, Olum 2002 for a defence of the SIA against the presumptuous philoso-

to prefer one over the other with near certainty just on grounds that it postulates vastly more observers. This, however, is exactly what the principle MS bluntly recommends. By way of contrast, the principle MU recommends $cr(T_1)/cr(T_2) = Pr(T_1)/Pr(T_2)$ (i.e. non-anthropic priors correspond directly to the rational credences), which seems much more plausible when applied to the Presumptuous Philosopher scenario.

To sum up the present investigation until now: The simple principle MU is natural and attractive at first sight, but it runs into trouble in Serpent's Advice and leads to the Doomsday Argument. The alternative principle MS avoids these problems but entails the SIA and, therefore, leads to the disastrous conclusion of the presumptuous philosopher. Arguably, what we need is a principle that coincides with MU in the Serpent's Advice scenario and the Doomsday Argument and with MS in the Presumptuous Philosopher scenario, ideally without radically changing the Bayesian framework that is used.[3] In the following section I propose a principle that seems to fulfil these expectations.

5 A causal principle

What might be the relevant difference between the Doomsday Argument and Serpent's Advice on the one hand and the Presumptuous Philosopher scenario on the other, so that MS seems to give the correct answer in the former and MU in the latter?

I suspect that the origin of our motivation to treat the two differently lies along the following lines: in the Presumptuous Philosopher scenario, we compare two possible worlds (corresponding to the two theories T_1 and T_2) that are "ready-made" in the sense that no conceivable action might contribute to make one rather than the other one actual. Thus, the credences $cr(T_1)$ and $cr(T_2)$ are about which one is and always has been actual. In Serpent's Advice, in contrast, Adam and Eve embark on considerations as to which world theirs is most likely to become, partly influenced by their actions. For them, whether H_1 or H_2 becomes true (i.e. whether they will have any offspring) depends causally on the actions they perform (or don't perform). Similarly, for humans in the Doomsday Argument who live before the asteroid strikes (or fails to strike), whether H_1 or H_2 holds is a question of which world theirs is to become in the future.

Let us consider the implications of these considerations for observers

pher challenge and Bostrom and Ćirković 2003 for a (to my mind convincing) rebuttal. Leitgeb 2010 offers an illuminating alternative version of the Presumptuous Philosopher scenario that compares hypotheses which differ on the number of observers by virtue of differing on the temporal extent of the universe (more specifically, on the number of expansion/contraction processes that the universe goes through).

[3]See, for instance, Bradley 2011 for good reasons not to abandon Bayesian conditionalisation in self-locating contexts.

whose self-locating information is maximally unspecific, e.g., in the Doomsday Argument, observers whose self-locating evidence is $H_{1,1} \vee ... \vee H_{1,N_1} \vee H_{2,1} \vee ... \vee H_{2,N_2}$. The considerations just offered suggest that observers with such maximally unspecific self-locating information are much more seriously disoriented in the Doomsday Argument and in Serpent's Advice than in the Presumptuous Philosopher scenario: in the former, they are unaware of whether they are among those who can potentially influence whether H_1 or H_2 is true or, in contrast, among those whose existence depends on H_2 being true. In particular, they are unaware of whether they are among those whose existence is potentially dependent on actions that, given who they might as well be, they themselves might in principle be able to perform. In contrast, observers in the Presumptuous Philosopher scenario whose self-locating information is maximally unspecific are not similarly disoriented because T_1 and T_2 are competing cosmological theories of which one is true and the other one false once and for all, such that there is no way of how any observer might conceivably influence whether T_1 or T_2 holds.[4]

Now to the central suggestion to be made in this paper. The main idea is the following: typically (most of the time in our everyday lives), observers can take over the non-anthropic priors as their rational credences; exceptions occur whenever they are disoriented about their place in the causal order of observers, in which cases they have to correct for effects that result from their being so disoriented. Let us try to make this suggestion formally precise inasmuch as it admits being made so.[5]

Let us introduce variables V_α, W_β, X_γ, ... (with Greek letters for the indices throughout) to denote complete collections of centres V_i, V_j, V_k, ..., and similarly for W, X, etc., between which no directed (direct or indirect) causal links exist. For collections V_α and V'_α from the same uncentred world V there are no constraints or limitations on joint members. I assume that whenever there are directed causal links between two centres V_i and V_j from the same world it is at least in principle possible either for V_i to prevent V_j

[4]One can of course modify the Presumptuous Philosopher problem and set it up such that, by assumption, it *becomes* at least in principle possible for observers to causally influence whether T_1 or T_2 holds. As I argue below, a good claim can be made that in such versions of the problem the presumptuous philosopher's MS-style reasoning is no longer presumptuous at all and the correct solution is analogous to the one in the Doomsday Argument.

[5]All present talk involving such notions as "causal structure", "causal order", "causal context" and "causal links" is meant to be understood as metaphysically non-committing, in particular not as presupposing causal realism. If one accepts a non-realist view of causation such as, for example, Huw Price's *causal perspectivalism* (Price 2007), one will have to interpret the terminology accordingly, for example, by tying causal relations and structure to the causal perspective that happens to be adopted by the epistemic agent. Making epistemic rationality perspectival in this sense might be a surprising move, but there is no reason why it should be in principle problematic. (Thanks to an anonymous referee for highlighting this implication of causal perspectivalism as applied to the present proposal)

from existing or for V_j to prevent V_i from existing.[6]

In theory, it may often be difficult to decide whether for some pair of centres V_i and V_j there are directed (direct or indirect) causal links between them, for example if candidate links exist that are for some contingent reason unexploitable for manipulative purposes. In such cases, there is no clear-cut answer to the question of whether there is some V_α to which V_i and V_j jointly belong. As a consequence, in these cases recommendations based on the principle to be formulated will be ambiguous and depend on the chosen interpretation of the causal relation between V_i and V_j.

In practice, however, determining whether there are causal links between any two centres V_i and V_j is typically straightforward. For instance, in Serpent's Advice, Adam is causally linked to both Eve and (indirectly) to his descendants (if there are any). As a consequence, the observers "Adam in H_1" and "Adam in H_2" are each individually already of the form V_α. In the Presumptuous Philosopher scenario, an example of a collection V_α is one which includes the presumptuous philosopher himself as well as a number of spacelike separated distant observers who are all causally isolated both from him and from each other. Worlds V which do not contain any observers are by definition themselves of the form V_α.

For the sake of simplicity I restrict attention to scenarios where within each possible world V the number of centres is the same within each collection V_α, V_β, V_γ, ... from the same uncentred V. This suffices to address (simple versions of) the scenarios discussed in the literature. In the absence of this assumption, the principle I am going to propose is no longer consistent with GenIND, so either it or GenIND will have to go in settings where it fails. Fortunately, in all the scenarios mentioned and in those to be discussed the assumption that all V_α, V_β, V_γ, ... for the same V have the same number of members seems viable. (Dropping it is a natural next step to be taken in future work, as the concluding section will argue.)

After these preliminaries, here is the principle I propose ("CP" stands for "causal principle"):

$$\text{(CP)} \qquad cr(V|V_\alpha \vee W_\beta \vee X_\gamma \vee ...) = Pr(V|V \vee W \vee X \vee ...)\,.$$

The disjunction $V_\alpha \vee W_\beta \vee X_\gamma \vee ...$ includes collections V_α, W_β, X_γ, ... of causally isolated observers from distinct worlds V, W, X, ... If we include a different collection α' from one world, say V, we obtain a new instance $cr(V|V'_\alpha \vee W_\beta \vee X_\gamma \vee ...) = Pr(V|V \vee W \vee X \vee ...)$. Similarly, if we simultaneously and/or alternatively replace β by β' and γ by γ', we arrive at further instances of CP.

[6]Strictly speaking, V_i cannot prevent V_j from existing as a matter of metaphysical necessity because both are centres within the same possible world V. The centre V_i may well possess causal powers to prevent V_j from existing, it may just not exercise them on pain of ceasing to be V_i.

To understand the ramifications of CP, it is useful to compare it with MU and MS. Crucially, CP reduces to MU if all centres in each world V, W, X, ... are causally isolated from each other, for in that case $V_\alpha = V$, $W_\beta = W$, $X_\gamma = X$ etc. On the other hand, CP reduces to MS if all centres in each world lie along one and the same directed chain of causal links between observers, for in that case $V_i = V_\alpha$ for each world and each i. The next section traces the ramifications of CP when applied to the problems of self-locating belief discussed above by highlighting these connections with MU and MS.

6 Applications

The principle CP is designed to reproduce the recommendation of MS in the Doomsday Argument and in Serpent's Advice and that of MU in the Presumptuous Philosopher scenario. This section investigates whether and, if so, how it does so.

Uncontroversially, in the Serpent's Advice scenario, directed (direct or indirect) causal links exist between Adam and Eve on the one hand and their descendants, i.e. all the later observers. In accordance with the remarks made at the end of the previous section, this reproduces the recommendation derived from MS: Adam and Eve may not legitimately assume that the risk of Eve's getting pregnant is negligible, so Adam's rational credences, knowing that he is the first observer ever to have existed, are $cr(H_1|H_{1,1} \vee H_{2,1}) = Pr(H_1)$ and $cr(H_2|H_{1,1} \vee H_{2,1}) = Pr(H_2)$.

Similarly, observers in the Doomsday scenario whose birth rank n it consistent with both H_1 and H_2 may in principle causally influence later observers, so their rational credences, according to CP, are $cr(H_1|H_{1,n} \vee H_{2,n}) = Pr(H_1)$ and $cr(H_2|H_{1,n} \vee H_{2,n}) = Pr(H_2)$. So, according to CP the Doomsday Argument fails.

Evidently, the recommendation that we wish to receive from CP for the Presumptuous Philosopher scenario is that the presumptuous philosopher's attitude is indeed presumptuous and that MU, rather than MS, leads to the right result here. However, the application of CP is not straightforward here since the story, as told by Bostrom or Leitgeb (see fn. 2), does not come with a specification of the causal links that exist according to T_1 and T_2, respectively.

To extract some recommendation from CP, let us stipulate that there is one civilisation of observers according to T_1 and a trillion of qualitatively identical copies of that civilisation according to T_2 in such a way that all those copies are causally isolated from each other. There are different ways of how causal isolation can arise: for example, it can be due to the fact that the spatial distances between civilisations are so large that causal influences, travelling no faster than the velocity of light, cannot possibly connect them;

or it can be due to the fact that, as in Leitgeb's modified Presumptuous Philosopher scenario (Leitgeb 2010), different civilisations occur in subsequent expansion and contraction cycles of an ever expanding and contracting cyclic universe, where causal influences are cut off (or are unexploitable for practical purposes as a matter of principle) by the periodically occurring big crunch/big bang-stages.

Considering this scenario, let $T_{1,\alpha}$ and $T_{2,\beta}$ be collections of observes that exist according to T_1 or T_2, respectively, whose members are all causally isolated from each other. As a consequence of the assumptions just made, the collection $T_{2,\beta}$ contains one trillion times more observers than $T_{1,\alpha}$. Applying the principle CP yields

$$cr(T_1|T_{1,\alpha} \vee T_{2,\beta}) = Pr(T_1)\,, \tag{6}$$

$$cr(T_2|T_{1,\alpha} \vee T_{2,\beta}) = Pr(T_2)\,. \tag{7}$$

By assumption, the number of observers in $T_{1,\alpha}$ differs from that of those who exist according to T_1 by the same factor as the number of observers in $T_{2,\beta}$ differs from that of those who exist according to T_2. As a consequence, we obtain

$$\frac{cr(T_1|T_1 \vee T_2)}{cr(T_2|T_1 \vee T_2)} = \frac{Pr(T_1)}{Pr(T_2)}\,, \tag{8}$$

in conformity with our intuitions and with MU, in contrast with the implausible result derived from MS (or the SIA).

But what are the recommendations of CP with respect to the Presumptuous Philosopher scenario if we allow (direct or indirect) causal links between the different civilisations that exist according to T_2? According to CP the existence of such causal links may have a nontrivial influence on the rational credences, and—one may ask—shouldn't the existence of such causal links be completely irrelevant for the rational credences?

In scenarios where the observers that exist according to T_1 and T_2 are causally linked among each other in complicated ways it is very difficult to check the plausibility of the recommendations given by CP, for we will typically not have any clear intuitions about the rational credences. However, there is an illuminating and important example that should be discussed, namely, where the class of observers that exist according to T_1 corresponds one-to-one to a subclass of the class of observers that exist according to T_2, while the vast majority of observers that exist according to T_2 (all those that are not in the subclass) are causal descendants (or, in an alternative scenario, ancestors) of those in the subclass. This constellation is realised, for instance, in a scenario where essentially all observers are causally linked among each other and T_1 and T_2 agree on the history of the world up to (or after) some time t, such that according to T_1 there are no observers after

(before) t, whereas, according to T_2, the vast majority of observers live after (before) t, all of them causal descendants (ancestors) of those who live before (after) t.

Once the Presumptuous Philosopher scenario is set up in this way, the recommendation given by CP is no longer the same as before. In fact, CP now recommends that all those observers who live before (after) t and are aware of it should align their rational credences with the non-anthropic priors, whereas those who are unaware of whether they live before or after t should prefer the theory T_2, which predicts a larger total number of observers, as if based on MS. Thus, with respect to this scenario, where T_1 and T_2 agree on the history before (after) t and disagree on it afterwards (before), the principle CP endorses the presumptuous philosopher's recommendation of the theory T_2 that predicts more obervers.

Luckily, this result, far from being worrisome for CP, speaks in favour of its plausibility. This is easily seen by noting that the scenario just described is essentially that of the Doomsday Argument: there as well the hypotheses at issue agree on the history of humanity up to some time t, one of them (H_1) predicts the extinction of humanity at t, whereas the other (H_2) predicts many more observers. Introducing causal links between the observers that exist according to T_2 along the lines just discussed has taken us from the original Presumptuous Philosopher problem to the scenario of the Doomsday Argument. The principle CP advises us to treat these cases differently, which—given our intuitive verdicts that the Doomsday Argument fails and the presumptuous philosopher's reasoning is presumptuous—is a virtue of CP, not a problem for it.

7 Sleeping Beauty

No other problem of self-locating belief has provoked the same amount of activity among epistemologists as the Sleeping Beauty problem. In its canonical exposition due to Adam Elga it reads as follows:

> Some researchers are going to put you to sleep. During the two days that your sleep will last, they will briefly wake you up either once or twice, depending on the toss of a fair coin (Heads: once; Tails: twice). After each waking, they will put you to [sic] back to sleep with a drug that makes you forget that waking. When you are first awakened, to what degree ought you believe that the outcome of the coin toss is Heads? (Elga 2000, 143)

Opinions are split over the correct answer. The two candidate rational credences for Beauty ("you", in Elga's example) with respect to "Heads" are 1/2 and 1/3, both of which have substantial support in the literature. The divide over the correct answer to the Sleeping Beauty problem is reflected

in the fact that the principles MU and MS produce exactly those diverging answers, and it is interesting to see which recommendation is given by CP.

To begin with, denote the three possible observer-moments (commonly called HEADS-Monday, TAILS-Monday, and TAILS-Tuesday) by H_1, T_1 and T_2, and assume that the coin toss is indeed fair in that $Pr(H) = Pr(T) = 1/2$ for the (objective) outcome probabilities. It is natural to use these probabilities as the non-anthropic priors. The self-locating evidence Beauty has at her disposal when awakening is maximally unspecific, i.e. $H_1 \vee T_1 \vee T_2$, which results in her rational credence

$$cr(H|H_1 \vee T_1 \vee T_2) = Pr(H) = 1/2 \tag{9}$$

according to MU and in

$$cr(H|H_1 \vee T_1 \vee T_2) = 1/2\,\frac{Pr(H)}{Pr(T)}\,cr(T|H_1 \vee T_1 \vee T_2) = 1/3 \tag{10}$$

according to MS (where the consequence of MS Eq. (5) has been used together with $cr(H|H_1 \vee T_1 \vee T_2) + cr(T|H_1 \vee T_1 \vee T_2) = 1$). Thus, the divide between MU and MS corresponds neatly to that between "halfers" and "thirders" on Sleeping Beauty.

Let us first consider the standard version of the problem where the two observer-moments T_1 and T_2 are treated as distinct stages of one and the same observer on two subsequent days (Monday and Tuesday). There are plenty of causal links from T_1 to T_2. For instance, if Beauty decides to have her hair dyed on Monday, she will, as a causal effect, have dyed hair if and when she is woken on Tuesday. It is true indeed that, in the story as told by Elga, Beauty on Monday has no practical means of affecting whether she is woken again on Tuesday. Nevertheless, we can imagine the experimenters asking her on any day—as a routine question—whether they may carry on with the experiment and wake her again (that is, if it is Monday and the coin falls TAILS), or whether, say, she feels sick and prefers that the experiment be aborted. Introducing such a question does not do violence to the story or transform it into one that calls for a totally different assessment. No such question could be asked if there were no causal links between T_1 and T_2 at all, in which case the Tuesday awakening could not possibly be prevented (or somehow influenced) on Monday.

Given the existence of causal links from T_1 to T_2, the principle CP coincides in its recommendation with MS:

$$cr(H|H_1 \vee T_1) = cr(H|H_1 \vee T_2) = Pr(H) = 1/2\,, \tag{11}$$

which, when combined with GenIND, yields the thirder answer

$$cr(H|H_1 \vee T_1 \vee T_2) = 1/3\,. \tag{12}$$

This result is attractive: even though the 1/2-answer has some notable proponents (such as Lewis 2001 and Bradley 2012), the 1/3-answer seems to remain the dominant one in the literature and, as it seems to me, for good reasons. Among the main independent arguments in its favour are the following two that go back to Elga: first, if the experiment is repeated many times, approximately 1/3 of the awakenings are HEADS-awakenings; second, on the 1/2-view, if Beauty is told that it is Monday, her rational credence with respect to HEADS turns to 2/3 by conditionalization.[7] As it does not seem to matter whether the coin is tossed on Sunday or Monday evening, this means that Beauty, fully oriented about her (temporal) location, should assign 2/3 to the outcome of a coin toss *yet to be tossed* being HEADS, which differs from its objective chance and seems difficult to accept.[8]

The most important objection against the 1/3-view is that it is difficult to see how its supporters can avoid commitment to the SIA and, thereby, however grudgingly, to the presumptuous philosopher's disastrous reasoning. The approach based on CP avoids this problem in that, as discussed in the previous section, its answer to the Presumptuous Philosopher problem (modulo complications that arise from the presence/absence of causal links) is that the presumptuous philosopher is indeed presumptuous, that is, as if based on MU (halfer-style reasoning). Thus, grounding thirdism on CP rather than MS avoids the only major significant challenge to being a thirder about the standard version of Sleeping Beauty.

Besides the standard version of the Sleeping Beauty problem just considered there is also a *fission* version, where if TAILS comes up, Beauty undergoes fission into two simultaneous observer-moments T_1 and T_2. In the fission scenario, if both are indeed created in in the same joint fission procedure, it is hard to see how any of the two post-fission Beautys could prevent the other one from being created in the first place. (One might immediately kill the other after the fission procedure has been performed, but then both will have coexisted at least for a very short while.) If this is the correct take on the fission version, the principle CP recommends the halfer answer in it by dictating $cr(H|H_1 \vee T_1 \vee T_2) = Pr(H) = 1/2$. Thus, on an interpretation of the fission version where the two post-fission Beautys cannot causally influence each other (at least not when starting to exist) the

[7]Some halfers, the so-called *double-halfers* (e.g. Bostrom 2007, Meacham 2008, Cozic 2011) dispute this. Meacham's position as espoused in (Meacham 2008) is especially interesting because he arrives at it on the basis of a proposed general principle of diachronic self-locating belief called compartmentalized conditionalization. This framework, however, is confronted with serious difficulties (see, for instance, (Titelbaum 2008, Section 2.7) and (Bradley 2011, Section 9), and in (Meacham 2010) he has abandoned it.

[8]See (Elga 2000) for these arguments and (Lewis 2001) for David Lewis' response supporting the 1/2-answer; for a condensed systematic overview of the debate about Sleeping Beauty, see (Titelbaum 2013); for an important recent defence of the 1/3-answer based on a general Bayesian framework that applies to self-locating belief, see (Titelbaum 2013, Chapter 9).

principle CP treats the temporal and the fission versions differently.

Is this difference in how CP treats the standard, "temporal", version of Sleeping Beauty on the one hand and the fission version on the other hand plausible? There are indeed important disanalogies between the temporal and the fission version: for example, while there is only one copy of Beauty in the temporal version, woken twice if TAILS comes up, there are two copies of her in the fission version if TAILS comes up, each woken only once. As a consequence, there is no obvious analogue to the long-run frequency argument in the fission scenario because it is unclear which of the persons that are woken in different runs of the experiment are the same. Furthermore, in the fission scenario it is not an option to perform the decisive coin toss on Monday evening, for at that time the fission procedure must have been made if it is to be made. As a consequence, Elga's second argument—which relies on the possibility of the coin being tossed on Monday evening—does not carry over to the fission scenario as well and there is considerably less motivation for thirdism in the fission version from the start.[9]

8 The Everett interpretation

The standard stochastic interpretation of quantum theory (SI) teaches that in each case where a measurement is performed exactly one of the possible outcomes is realised. The probabilities for the various possible outcomes are given by their *Born weights*, which in turn are computable by the formalism of quantum theory. In contrast to the SI, the Everett interpretation of quantum theory (EI) teaches that in each case where a measurement is performed *all* possible outcomes of the measurement are realised, but in different *branches* of a vast Everettian "multiverse", each branch associated with its distinctive Born weight.

Since the EI is designed to reproduce the empirical predictions of the SI in all but very special experimental circumstances that are extremely hard to bring about, one would expect that evidence which confirms (or disconfirms) the SI typically also confirms (or disconfirms) the EI to the same degree. There are various issues about how empirical confirmation is accommodated in the EI, and here I focus only on the most dramatic worry, namely, that, since the EI predicts that *all* possible outcome really occur, any arbitrary outcome confirms it, in contrast with the SI, which makes very specific predictions. If this were true, it would mean that, as noted by Bradley, "[t]he Ancients could have worked out that they have overwhelming

[9]See (Kierland and Monton 2005) for supporting considerations, using two competing criteria of expected inaccuracy, that the temporal and the fission versions might best be assessed differently and Wilson 2014 for support based on principled considerations about the relation between chance and rational credence. Some diachronic approaches Meacham 2010, Schwarz 2012, 2015 arrive at the conclusion that the temporal and fission versions require different treatments along entirely independent lines.

evidence for [EI] merely by realizing it was a logical possibility and observing the weather." (Bradley 2012, 159) Clearly the case for the EI cannot possibly be as simple as that, and the reasoning that leads to trivial confirmation of the EI must be somehow fallacious.

However, as pointed out in (Lewis 2007), while the naive confirmation view of the EI is obviously inadequate, it seems as if endorsing the 1/3-view on Sleeping Beauty commits one to it.[10] Put as simply as possible, the argument for this claim is this: according to the 1/3-view on Sleeping Beauty, it is rational for Beauty to prefer TAILS over HEADS when awaking, related to the fact that there are more awakenings in the TAILS-world than in the HEADS-world. Applied to the dispute between the SI and the EI, an analogous line of thought leads to the result that we should prefer the EI over the SI, just on grounds that it predicts more observer-moments, corresponding to all possible measurement outcomes, which according to the EI are in fact all *real* measurement outcomes. Fortunately, the principle CP provides a rationale for not drawing this unattractive conclusion and for retaining thirdism while rejecting naive confirmation of EI.

To apply CP to the SI/EI-dispute in a simple example, let us consider the measurement setup studied in (Lewis 2007) and in (Bradley 2012): a spin-1/2 particle prepared in a state where the Born weights for the possible measurement results $+1/2$ and $-1/2$ of, say, spin in z-direction, are identical, i.e. $BW(+1/2) = BW(-1/2) = 1/2$. There are four possible observer-moments to be considered, namely EI_+, EI_-, SI_+ and SI_-, which can be read as "I am in the Everett branch where the outcome is $+1/2$", "I am in the Everett branch where the outcome is $-1/2$", "The SI holds and the outcome is $+1/2$", and "The SI holds and the outcome is $-1/2$", respectively. There are no causal links between any two of these observer-moments: occurrences of SI_+ and SI_- are mutually exclusive, and both are incompatible with EI_+ and EI_-. The latter two correspond to distinct branches of reality, and it is impossible for an observer detecting the outcome $+1/2$ in the "up-branch" to influence the situation of an observer detecting the outcome $-1/2$ in the "down-branch" (and conversely). In particular, whether or not, say, the observer-moment EI_- exists at all is not something that the observer-moment EI_+ can causally influence.[11]

[10] Bradley endorses Peter Lewis' reasoning together with its conclusion, arguing that it is to be welcomed by the Everettian (Bradley 2012, Section 2.1). Wilson disputes that Everettians should happily embrace the 1/2-view but also argues that they may escape committing themselves to it by relying a principled distinction between chancy and ignorance-based input probabilities (Wilson 2014). This is an interesting proposal, which, in the light of the close conceptual ties between chance and causation, is perhaps not unrelated to the present one. See (Bradley 2015) for criticism that it would need to overcome.

[11] According to some versions of the Everett interpretation, split branches may in principle recombine. Cross-branch causal influences may occur in these cases and, if we follow CP, may have a non-trivial impact on the rational credences in more complicated situa-

If we plug in the observation that EI_+, EI_-, SI_+, SI_- are not causally linked and denote our non-anthropic prior assigned to the Everett interpretation (on whatever systematic and/or interpretive grounds we have arrived at it) by $Pr(EI)$, the principle CP agrees with the principle MU in that it yields

$$cr(EI|EI \vee SI) = cr(EI|EI_+ \vee EI_- \vee SI_+ \vee SI_-) = Pr(EI)\,. \qquad (13)$$

To see the consequences of this result, assume that a measurement of spin in z-direction is performed, and the outcome is, say, $+1/2$. As emphasised by Bradley, it is crucial to note here that the evidence gained by the experimentalist is not merely "There exists a $+1/2$-branch" (which would correspond to $EI_+ \vee EI_- \vee SI_+$), but the more specific "I am in a $+1/2$-branch" (which corresponds to $EI_+ \vee SI_+$). Thus, the experimentalist's rational credence after having registered the result $+1/2$ is $cr(EI|EI_+ \vee SI_+)$.

The crucial assumption made by the Everettians is that the rational credences of being in the up- or down-branch are given by the Born weights of the outcomes $+1/2$ and $-1/2$. In other words, Everettians assume that the rational credences are $cr(EI + |EI) = BW(+1/2)$ and $cr(EI - |EI) = BW(-1/2)$, where $BW(+1/2)$ and $BW(-1/2)$ are the corresponding Born weights. *Given* this assumption (it is highly controversial whether Everettians are justified in availing themselves of it), the following result can be derived by using CP's instance Eq.Â (13):

The crucial assumption made by the Everettians is that the rational credences of being in the up- or down-branch are given by the Born weights of the outcomes $+1/2$ and $-1/2$. In other words, Everettians assume that the rational credences are $cr(EI_+|EI) = BW(+1/2)$ and $cr(EI_-|EI) = BW(-1/2)$, where $BW(+1/2)$ and $BW(-1/2)$ are the corresponding Born weights.[12] *Given* this assumption (of which it is highly controversial whether Everettians are justified in making it), the following result can be derived by using CP's instance Eq. (13):

$$\begin{aligned} cr(EI|EI_+ \vee SI_+) &= \frac{cr(EI_+ \vee SI_+|EI) \cdot cr(EI)}{cr(EI_+ \vee SI_+)} \\ &= \frac{BW(+1/2) \cdot Pr(EI)}{BW(+1/2)} \\ &= Pr(EI)\,. \end{aligned} \qquad (14)$$

Bayes' Theorem has been used in the first line and Eq. (13) in the second line. In addition, it has been assumed that no further interpretations besides SI

tions.

[12] According to Vaidman 1998, this is an instance of the more general basic (not further justifiable) Everettian principle that rational credences should be assigned in conformity with the Born weights conceived of as *measures of existence*. According to Wallace 2012 (Part II in particular), in contrast, decision-theoretic considerations can be used to provide deeper justifications of this use of the Born weights.

and EI need to be considered: $cr(EI) = cr(EI|EI \vee SI)$. The result Eq. (14) is reassuring. It means that evidence that the outcome is $+1/2$ (or $-1/2$) does *not* automatically lead to confirmation of the Everett interpretation over the stochastic interpretation, exactly as it should.

As indicated, the remaining challenge for the Everettians is to justify that the rational credence of being in some branch really coincides with the associated Born weight. In the example considered, Everettians must justify $cr(EI_+|EI) = BW(+1/2)$ and $cr(EI_-|EI) = BW(-1/2)$ as the rational credences for *arbitrary* Born weights $BW(+1/2)$ and $BW(-1/2)$. Matters are highly subtle here. To highlight just one of the complications that arise, consider a scenario where only the observer who witness the result $+1/2$ survives, i.e. where the only possible (post-measurement) observer-moments are SI_+ and EI_+.[13] In the formalism used here EI is decomposed as $EI = EI_+ \vee EI_-$ if both EI_+ and EI_- exist. However, if there is no observer-moment EI_-, the decomposition reduces to $EI = EI_+$, which yields $cr(EI_+|EI) = 1$, differing from the Born weight $BW(+1/2) \neq 1$. If Everettians are committed to this conclusion, their interpretation is empirically inadequate, which is bad news for them. In response to this difficulty, Everettians might opt for amending the formalism by adding *ghost* observer-moments in unpopulated branches such as EI_- even if they do not really exist. It seems difficult to judge whether this move can be made in a non-ad hoc way without delving much more profoundly into how the Everett interpretation should be fleshed out in detail.

On the more optimistic side, however, let us recapitulate the main moral of this section, which is that CP allows to combine thirdism about the temporal version of Sleeping Beauty with a view according to which the Everett interpretation is not automatically confirmed by arbitrary empirical data. The additional question of whether the Everett interpretation is empirically equivalent with the SI whenever it should be—including less simple examples than the one considered here—is much more difficult and beyond the scope of this paper.

9 Concluding remarks

I have motivated and defended CP as a promising recipe for obtaining the rational credences from the non-anthropic priors in the scenarios considered. My main goal, however, is *not* to establish CP as the ultimately correct way to to this—due to its in-built limitations and vagueness it almost certainly isn't—but to direct our attention to the *causal* contexts of observers as potentially relevant for the correct solutions to problems of self-locating belief.

[13] An anonymous reviewer points out Lewis 2004 as an influential development of this objection.

Future work that takes up this recommendation could elaborate on the present discussion in at least the three ways: first, it may try to find a version of CP that applies to scenarios where indifference principles like IND and GenIND are implausible from the start—for example scenarios where collections V_α and $V_{\alpha'}$ that belong to the same uncentred V differ in the number of members. Second, they may try to put CP (or a refined version) on a firmer philosophical basis, for example by recovering it in the language of causal graphs (Pearl 2000, Spirtes et al. 1993) or by investigating it in the language of the Principal Principle (Lewis 1986).[14] Third, they may want to explore the consequences of CP (or a refined version) when applied systematically to a wider class of problems of self-locating belief, appropriately categorized, including, for example, the problem of empirical confirmation in a multiverse where the constants of nature are different in the various subuniverses.[15]

Acknowledgements

I would like to thank various anonymous referees and the participants of my research seminar on self-locating belief at Göttingen University in the winter term 2012/13, where I developed the ideas presented here. Furthermore, I would like to thank audience members in Groningen and Munich for their comments and questions.

References

Bostrom, Nick 2001: "The Doomsday Argument, Adam & Eve, UN^{++}, and Quantum Joe". *Synthese* 127, 359-387.

Bostrom, Nick 2002: *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. New York: Routledge.

Bostrom, Nick 2002: "Self-locating Belief in Big Worlds: Cosmology's Missing Link to Observation". *Journal of Philosophy* 99, 607-623.

Bostrom, Nick 2007: "Sleeping Beauty and Self-location: A Hybrid Model". *Synthese* 157, 59-78.

[14] Phrased in this language, when the non-anthropic priors are Lewisian chances CP comes down to the statement that self-locating evidence is *admissible* just in case it informs one of one's place in the causal order of observers and *inadmissible* otherwise. Accordingly, a promising strategy to defend CP as applied to Lewisian chances is by arguing for this condition on the admissibility of self-locating evidence.

[15] Invoking a multiverse of this type is a standard response to the problem of apparent fine-tuning for human life of the constants of nature in our universe, see, for instance, McMullin 1993, Bostrom 2002b, Bradley 2009 for assessments of some of the intricacies that arise in a multiverse context from philosophical points of view.

Bostrom, Nick and Ćirković, Milan M. 2003: "The Doomsday Argument and the Self-indication Assumption: Reply to Olum". *The Philosophical Quarterly* 53, 83-91.

Bradley, Darren 2009: "Multiple Universes and Observation-selection Effects". *American Philosophical Quarterly* 46, 61-72.

Bradley, Darren 2011: "Self-location Is no Problem for Conditionalization". *Synthese* 182, 393-411.

Bradley, Darren 2012: "Four Problems of Self-locating Belief". *Philosophical Review* 121, 149-177.

Bradley, Darren 2015: "Everettian Confirmation and Sleeping Beauty: Reply to Wilson". *British Journal for the Philosophy of Science* 66, 683-693.

Briggs, Rachel 2010: "Putting a Value on Beauty". In: Tamar Szabo Gendler and John Hawthorne (eds.), *Oxford Studies in Epistemology, Volume 3*. Oxford: Oxford University Press, 3-34.

Carter, Brandon 1974: "Large Number Coincidences and the Anthropic Principle in Cosmology". In: Malcolm S. Longair (ed.), *Confrontation of Cosmological Theories with Data*. Dordrecht: Reidel, 291-298.

Cozic, Mikal 2011: "Imaging and Sleeping Beauty: A Case for Double-halfers". *International Journal of Approximate Reasoning* 52, 137-143.

Dieks, Dennis 2007: "Reasoning about the Future: Doom and Beauty". *Synthese* 156, 427-439.

Elga, Adam 2000: "Self-locating Belief and the Sleeping Beauty Problem". *Analysis* 60, 143-147.

Elga, Adam 2004: "Defeating Dr. Evil with Self-locating Belief". *Philosophy and Phenomenological Research* 69, 383-396.

Gott, J. Richard 1993: "Implications of the Copernican Principle for Our Future Prospects". *Nature* 363, 315-319.

Hawley, Patrick 2013: "Inertia, Optimism and Beauty". *Nous* 47, 85-103.

Kierland, Brian and Monton, Bradley 2005: "Minimizing Inaccuracy for Self-locating Beliefs". *Philosophy and Phenomenological Research* 70, 384-395.

Kim, Namjoong 2009: "Sleeping Beauty and Shifted Jeffrey Conditionalization". *Synthese* 168, 295-312.

Leitgeb, Hannes 2010: "Sleeping Beauty and Eternal Recurrence". *Analysis* 70, 203-205.

Leslie, John 1996: *The End of the World: The Science and Ethics of Human Extinction*. London: Routledge.

Lewis, David 1979: "Attitudes De Dicto and De Se". *The Philosophical Review* 88, 513-543.

Lewis, David 1986: "A Subjectivists's Guide to Objective Chance". In: *Philosophical Papers, Vol. II*. New York: Oxford University Press, 83-132; originally published 1980 in: Richard C. Jeffrey (ed.) *Studies in Inductive Logic and Probability, Vol. II*. Berkeley: University of California Press.

Lewis, David 2001: "Sleeping Beauty: Reply to Elga". *Analysis* 61, 171-176.

Lewis, David 2004: "How Many Lives Has Schrödinger's Cat?". *Australasian Journal of Philosophy* 82, 3-22.

Lewis, Peter J. 2007: "Quantum Sleeping Beauty". *Analysis* 67, 59-65.

Lewis, Peter J. 2010: "A Note on the Doomsday Argument". *Analysis* 70, 27-30.

McMullin, Ernan 1993: "Indifference Principle and Anthropic Principle in Cosmology". *Studies in History and Philosophy of Science* 24, 359-389.

Meacham, Christopher J. G. 2008: "Sleeping Beauty and the Dynamics of *De Se* Belief". *Philosophical Studies* 138, 245-269.

Meacham, Christopher J. G. 2010: "Unravelling the Tangled Web: Continuity, Internalism, Uniqueness, and Self-locating Belief". In: Tamar Szabo Gendler and John Hawthorne (eds.), *Oxford Studies in Epistemology, Volume 3*. Oxford: Oxford Universit Press, 86-125.

Moss, Sarah 2012: "Updating as Communication". *Philosophy and Phenomenological Research*, 85, 225-248.

Neal, Radford M. 2006: "Puzzles of Anthropic Reasoning Resolved Using Full Non-indexical Conditioning". `http://arxiv.org/abs/math/0608592`.

Norton, John 2010: "Cosmic Confusion: Not Supporting versus Supporting Not-". *Philosophy of Science* 77, 501-23.

Olum, Ken 2002: "The Doomsday Argument and the Number of Possible Observers". *The Philosophical Quarterly* 52, 164-184.

Pearl, Judea 2000: *Causality*. New York: Cambridge University Press.

Pisaturo, Ronald 2009: "Past Longevity as Evidence for the Future". *Philosophy of Science* 76, 73-100.

Price, Huw 2007: "Causal Perspectivalism". In: Huw Price and Richard Corry (eds.), *Causation, Physics and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Oxford University Press, 250-292.

Schulz, Moritz 2010: "The Dynamics of Indexical Belief". *Erkenntnis* 72, 337-351.

Schwarz, Wolfgang 2012: "Changing Minds in a Changing World". *Philosophical Studies* 159, 219-239.

Schwarz, Wolfgang 2015: "Belief Update across Fission". *British Journal for the Philosophy of Science* 66, 659-682.

Stalnaker, Robert C. 2008: *Our Knowledge of the Internal World*. Oxford: Oxford University Press.

Spirtes, Peter, Glymour, Clark, and Scheines, Richard 1993: *Causation, Prediction and Search*. New York: Springer.

Titelbaum, Michael G. 2008: "The Relevance of Self-locating Beliefs". *Philosophical Review* 117, 555-605.

Titelbaum, Michael G. 2013: *Quitting Certainties: A Bayesian Framework Modelling Degrees of Belief*. Oxford: Oxford University Press.

Titelbaum, Michael G. 2013: "Ten Reasons to Care about the Sleeping Beauty Problem". *Philosophy Compass* 8, 1003-1017.

Vaidman, Lev 1998: "On Schizophrenic Experiences of the Neutron or Why We Should Believe in the Many-worlds Interpretation of Quantum Theory". *International Studies in the Philosophy of Science* 12, 245-261.

Vilenkin, Alexander 1995: "Predictions from Quantum Cosmology". *Physical Review Letters* 74, 846-849.

Wallace, David 2012: *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation*. Oxford: Oxford University Press.

Weatherson, Brian 2005: "Should We Respond to Evil with Indifference?". *Philosophy and Phenomenological Research* 70, 613-635.

Wilson, Alastair 2014: "Everettian Confirmation and Sleeping Beauty". *British Journal for the Philosophy of Science* 65, 573-598.